STELLA: A Scheme for a Learning Machine

J. H. ANDREAE

(Submitted 23 August 1962 to the 2nd IFAC Congress, held at Basel, Switzerland, in 1963 and published in the Conference Proceedings: *Automation & Remote Control*, ed. Broida, Butterworths (1969) pp.497-502.)

Summary

A scheme for a learning machine is described. In a basic exploratory mode the machine searches its environment. Learning enables it to profit from those action sequences which lead to reward. By correlating the changes in its environment with its actions, the machine can extract invariant features and use them to guess profitable actions when its learned sequences fail. An internal mode of operation is described in which the machine explores the possibilities of its future actions with a view to modifying its perforance. The learning automaton, STELLA, is being constructed in the form of a mechanical tortoise which takes its name from its laboratory origin: Standard Telecommunication Laboratories Ltd.

Introduction

STELLA is a scheme for a general-purpose learning machine. For convenience, the scheme is being tested in the form of a mechanical 'tortoise' moving about on a level floor which can observe the approximate distances and angles of obstructing walls. The scheme is intended to be unspecific to the particular type of environment in which the machine operates, and other machines and environments are under consideration.

There would appear to be two main approaches to the design of learning machines. In the one, a rigid control system performing a strictly determined task is modified to make it less rigid, more flexible and adaptive. In the other, a general-purpose learning machine is designed and is then modified to make it more specific to particular problems. STELLA falls within this latter category and the object of experimenting with the STELLA scheme is to explore ways in which it can evolve into a more specialized system. The importance of this process of 'progressive segregation' of parts is emphasized by Bertalanffy[1].

STELLA, the tortoise, is a self-propelling and self-steering trolley which can move about on a level floor. Walls of standard height having illuminated upper rims obstruct its movements in various directions. Information about the angular positions and distances of these walls is obtained by photocells mounted around the circumference of the trolley. There are eight of these photocells on the machine which is being constructed. *Figure 1* assumes six only.

The Exploratory Mode

Unless its memories are 'primed' by the experimenter, STELLA starts with no information about its environment. In these circumstances, and whenever the higher level of operation fails to decide its actions, STELLA proceeds to an exploratory or search mode. This mode is governed by a random generator in such a way that the machine tends to perform different actions when receiving the same information repeatedly from the photocells of its eye.

Each photocell of the eye has a fixed threshold which the light input to the photocell must exceed for the information from that photocell to change from -1 to +1.

Thus each photocell contributes one "bit" to the input binary pattern from the eye---the eye pattern.

The actions of STELLA are restricted, in *Figure 1* and in the constructed machine, to four movements: forward, forward to the left, forward to the right, and reverse.

The short-term memory shown in *Figure 1* consists of a 6 x 4 matrix of binary elements represented by the intersections of the six column lines from the eye and the four row lines to the dissimilarity detector (DD). Each row line is associated with one of the output movements, forward, left, right or reverse.

When the eye 'sees' a pattern, the DD selects that row in the short-term memory along which is stored the pattern least similar to the eye pattern. The motor output control then performs the output movement or action corresponding to the selected memory row. The continuous random improver of the memory switches, by random choice, some of those elements in the selected row which are not the same as the corresponding elements of the eye pattern, until, due either to the movement causing a change in the environment or to the random improvement of the row, that row ceases to be the least similar to the eye pattern and the DD selects a different row. The selection of least similar rows follows the procedures of Steinbuch[2]; the random improvement is superimposed to provide a short-term memory which eliminates repetitive reactions to the eye pattern: STELLA should not push persistently against walls and should break out of repetitive cycles.

Learning by Reward

Patterns and actions can be stored in the two parts of the long-term step memory shown in *Figure 1.* At the same time, and in the same row, a connection can be made in the sequence store to indicate that the pattern and action of that row was followed by a pattern and action stored in another row, or by reward. Each intersection between a row line and column line in the sequence store represents the following of the action of that row by the pattern in the row connected to that column line; or, if it is the reward column (marked 'pleasure' in *Figure 1*), the following of the action of that row by reward. There may be different kinds of reward with separate columns, and there may be a 'pain' column; these are considered in later sections.

When an eye pattern is received, it is compared with the patterns stored in the long-term step memory by the threshold detector (TD). The TD associates any row of the memory which has a pattern more similar to the eye pattern than the current threshold of similarity demands. Consider three cases:

(a) The TD associates no row, but the action prescribed by the DD leads to reward. The eye pattern and the subsequent action are stored in a vacant row of the long-term step memory and a connection is made in the sequence store between that row and the reward column. The row is given unit weight in the weight store.

(b) The TD associates one row. The TD overrides the DD and causes the motor output control to perform the action stored in the associated row of the long-term step memory. At the same time, suppose that there is a connection in the sequence store which indicates that on a previous occasion the action was followed by the pattern in another row of the

long-term step memory. The weight of this other row is increased in anticipation of its being seen again. Also, while the action is being performed, the preceding pattern and action are held tentatively in a vacant row of the long-term step memory; in fact, every pattern and action is held tentatively for one cycle. Now, the action performed, several situations may ensure:

(i) If there is no association of the new eye pattern by the TD and no reward, the weight of the anticipated row is reduced to what it was, the tentatively held pattern and action are disregarded, the weight of the last row to be associated by the TD is reduced, and the machine STELLA reverts to the search mode.

(ii) If reward is received, the weight of the associated row is increased, a connection in the sequence store is made to the reward column, and the tentatively held pattern and action are stored with unit weight in a row with a sequence connection to indicate the row which followed.

(iii) If the anticipated row is associated by the TD, its weight is increased further in confirmation, the tentatively held pattern and action are held for another cycle and the next step is entered.

(iv) If an unanticipated row is associated by the TD, the weight of the anticipated row is decreased and the weight of the associated row is increased.

(v) If the TD associates more than one row, one has, as for (c).

(c) The TD associates more than one row. Each of the associated rows has a weight depending upon how many times it has been used in the past and each row is connected to the reward column through a series of connections in the sequence store. Each connection in this 'circuit' of the sequence store between the row and the reward column represents a possible step. If each connection is imagined to have an electrical resistance and the column and row lines of the sequence store are imagined to be wires, then the total resistance between an associated row and the reward column is a measure of the number of steps expected to lead to reward. The random row selector chooses one of the rows associated by the TD on a chance basis biased towards high row weights and short paths to reward. Once a single row has been chosen, the situation is the same as for (b).

Figure 2 is a flow diagram of the processes outlined above, together with other processes which will be described.

The threshold of the TD determines the permitted latitude in the association of row patterns with the eye pattern. If the threshold of similarity is low, the behaviour will tend to be illogical and opportunity will exist for the discovery of shorter paths to reward. When STELLA is being particularly successful in obtaining rewards, it is necessary for the threshold to be kept high so that previously established paths will be followed more precisely the more more successful they prove to be. However, it is not desirable that a pattern should be completely disregarded in favour of search behaviour because the threshold happens to be *just* too high. Therefore, it is proposed to have a slow decline in the value of the threshold with time, to cause the threshold to be restored to a maximum when reward is received, and to allow the TD to depress its own threshold (temporarily) by not more than a predetermined amount below its current value, if no pattern is associated.

'Forgetting' is necessary to remove those steps which are less used and, therefore, less useful so as to make room for more important steps. To do this the weights of the rows in the long-term step memory are given slow time decays. When a weight falls below some prescribed value, the contents of the row are erased and all connections to and from the row in the sequence store are broken.

The 'teacher', shown in *Figure 1*, is a control which overrides the motor output control to enable the experimenter to speed up the learning process by forcing STELLA along profitable paths.

Pain

There will be some values of the parameters controlled by a learning sytem which must be avoided. For example, in the control of a chemical plant it may be known that certain mixtures of reactants are explosive and that these must be avoided; again, in a traffic control system it is likely that there will be a minimum safe distance of approach of vehicles. In the tortoise floor-wall system one can imagine holes in the floor which the tortoise must avoid in order to survive. Clearly it is inadequate to provide learning by reward only in a machine which can act by trial and error in its environment if forbidden conditions are to be avoided. It must be taught or be programmed to avoid these conditions and the information which determines this avoidance will be called pain. Responses to pain cannot be maintained by experience in the way responses to reward are reinforced. A traffic control sytem could not be allowed to maintain a safe distance of approach between vehicles by regular experiments with unsafe distances.

Pain is introduced into the STELLA scheme of *Figure 1* by the addition of a 'pain' column to the sequence store and by an externally controlled 'pain warning'. In order to understand the implication of the negative sign attached to the pain column, which may be compared with the positive sign ascribed to the pleasure, or reward column, it is convenient to think in terms of the actual electrical method employed for the operation of the random row selector. In order to bias the decision of this selector according to the length of the path (number of steps) to reward, the connections in the sequence store are resistors and the row line in question is earthed., while the reward column line is connected to a positive battery voltage. In this way a current flows from battery positive to the earthed row which is inversely proportional in magnitude to the resistance of the path, that is, the length of the path. Now some connections to the pain column will be programmed into the machine and some will be formed as a result of the pain warning. If the pain column is connected to a negative battery voltage, the currents to this column, in so far as they pass through the same resistors as the current from the positive reward column, will oppose the currents from the pleasure column and decrease the probability that the random row selector will choose the row in question. Actions expected to lead to pain will be inhibited according to how many steps are expected before pain may be encountered. If the pain currents exceed the positive currents, it is arranged for the action prescribed by the row to be blocked. That action is forbidden for that step.

The conenctions to the pain column are made in the same way as for those to the reward column, but allowance is made for preprogrammed connections to the pain column to be established for some rows with permanent (non-decaying) weight.

Specialization

In order that STELLA should learn paths out of painful situations, the cessation of the pain warning should be treated as a reward and an additional positive reward column is required for the sequential store. This column would become connected to the positive battery voltage when a pain warning was received and it is logical to arrange for the other reward column to be disconnected simultaneously. Suppose that the machine is given a goal and the pleasure reward indicates the achievement of this goal. If the implication of a pain warning is sufficiently serious, as it is intended to be, then the final attainment of the goal is subject to a temporary diversion to the subsidiary goal of escaping from the painful situation. Similarly, if the system has to ensure the supply of its own power or the supply of raw materials to the process it is controlling, it may be essential for this secondary goal to be attended to at the expense of progress toward the prime goal. This logical interdependence of the operation of the reward columns forces upon us the specialization of the system to match its environment. The machine must be told to begin with, what is required of it. It must also be given some rules to prevent it from destroying itself or others. Natural selection is too costly.

What other kinds of specialization have to be introduced? The scheme of *Figure 1* was designed to be unspecific to its environment. The connections from the eye can be interchanged so that different photocells control different columns of the memories and after a time the system should adapt itself to the new arrangement. The same applies to the connections to the output actions. But this interchangeability of connections presumes that each photocell is providing equivalent information; otherwise patterns stored in the long-term step memory will give undue importance to digits on some of the columns at the expense of those on others.

It seems possible for STELLA to compensate for the unequal importance of digits in the eye pattern. Each column is given a variable weight which determines the bias applied to the appropriate element of a pattern associated by the DD or TD. The column weight is decreased every time the TD disregards the corresponding digit of the eye pattern in associating a pattern in the long-term step memory with the eye pattern. After a time some columns will be contributing little to the operation of the machine and the experimenter can change the position or other property of the respective photocells in order to achieve an arrangement in which these photocells contribute their full share. The machine could, of course, be programmed to make such changes itself. However, the experimenter might find that more radical changes were needed. For example, it might be more effective to start with 100 photocells logically connected to give only the six outputs, these outputs representing more specific characteristics of the environment. See, for example, Lettvin *et al.*[2]

The variation in the column weights should represent the relative importance of the information arriving on the various columns, when the importance is averaged over a large number of steps, and it should indicate the efficiency of the photocells as receptors of information. The relative importance of elements of a pattern stored in the long-term step memory will depend, not only on the efficiency of the receptors, but also on the situations in which the pattern is used. Suppose that the experimenter decides to reward STELLA (the tortoise) every time it reverses on approaching a wall. As things have been

described, STELLA would have to store a number of patterns representing the various situations in which it approaches a wall, each of these patterns being coupled with the action 'reverse'. If the machine is not being very successful in obtaining reward and the threshold is low, then it may be lucky enough to use one of these patterns in a situation which does not quite correspond to the remembered situation, but this is not possible when it is more successful. Now let each element of each pattern in the long-term step memory have variable weight, the weights for each pattern being normalized so that the TD will still just associate this pattern with the eye pattern if it is identical and if the threshold is at a maximum. The element weights are modified each time the row is selected by the random row selector, disregarded elements having their weights lowered. The experiment envisaged above will lead to the establishment of a pattern in the long-term step memory coupled to the action 'reverse' with its element weights adjusted so that only those digits of the pattern, which indicate the presence of a wall ahead, contribute to the association procedure of the TD. A crude kind of generalization takes place so that a number of possible patterns representing a particular situation are accepted and remembered as a single pattern.

The Correlation Level

It was stated above that to begin with the machine must be told what is required of it, but this is not strictly true. It can explore its surroundings and learn something about their characteristics while the experimenter is still making up his mind about what he wants the machine to do. STELLA does this constantly by means of its correlation matrices, which are shown, drawn one upon the other, in *Figure 1*. Each of the four matrices is associated with one of the actions.

Every eye pattern received is applied to the columns of the matrix appropriate to the action which has just been performed. The same pattern is then held for one step (one action) applied to the rows of the matrix corresponding to that action. Thus, for any action, the appropriate matrix has the initial pattern applied to its rows and the resultant pattern applied to its columns. So long as the action has not been prevented by the external nevironment, each element of the matrix has its 'value' increased (decreased) if its row and column have the same (opposite) binary values (±1).

When STELLA has explored the environment for some time, the correlative matrices should contain experience of the way in which its actions transform the observed environment. Invariants in these transformations will be reinforced by persistent increases or decreases in the values of particular elements of the matrices. For example, in the tortoise floor-wall arrangement one would expect the matrix corresponding to the forward movement to contain information that an observed pattern will move past the tortoise from front to back.

The correlation matrices are used if the TD fails to associate a pattern when a remembered path is being followed. The sequence of operations is shown in *Figure 2*. The unassociated pattern is applied to each of the correlation matrices in turn until the TD associates one of the transformed patterns, then the action corresponding to the matrix which effected the transformation is performed, as a hopeful guess based on experience. The transformation by one of the matrices is carried out by matric multiplication of the

pattern binary vector by the non-binary matrix, the sign of the componenets of the product vector determining the binary elements of the transformed pattern.

If the guesses resulting from the correlation matrices sometimes enable STELLA to rediscover the paths which it loses, they will contribute to the efficiency and speed of learning. There is, however, a more significant way in which the correlation matrices can take part in predictive forecasts. This is the subject of the next section.

The Internal Mode

If the reception of eye patterns and the performance of actions are blocked, an internal mode of operation can be envisaged which might be called 'dreaming'. The last eye pattern to get through before the blockage leads by the DD or the TD to the selection of an action but, instead of the action being performed, the corresponding correlation matrix is used to transform the eye patter into a 'guessed' second pattern. This second pattern is now treated as a new eye pattern, the DD and TD select a correlation matrix to form a third pattern, and so on. The 'dreams' may lead sometimes through the random excursions of the DD and sometimes more logically by the TD through the remembered paths of the long-term step memory to occasions of pain and pleasure.

In the internal mode of operation, STELLA is exploring the possibilities of future actions by using the information stored in the correlation matrices and in the long-term step memory to anticipate the effects of its postulated actions. If time is allotted during the perormance of each action (see *Figure 2*) for short excursions in the internal mode, and if some variation of the weights of rows in the long-term memory is permitted when the excursions anticipate reward or pain, then STELLA can modify its own actions according to the machine's predictions.

References

[1] VON BERTALANFY, L. An outline of general systems theory. *Brit.J.Phil.Sci.* **1,** 2 (1950) 134.

[2] STEINBUCH, K. Die Lernmatrix. *Kybernetik* **1,** 1 (1951) 36.

[3] LETTVIN, J. Y., MATURANA, H. R., McCULLOCH, W. S. and PITTS, W. H. What the frog's eye tells the frog's brain. *Proc. Inst. Radio Engrs, N.Y.* **47,** 11 (1959) 1940.

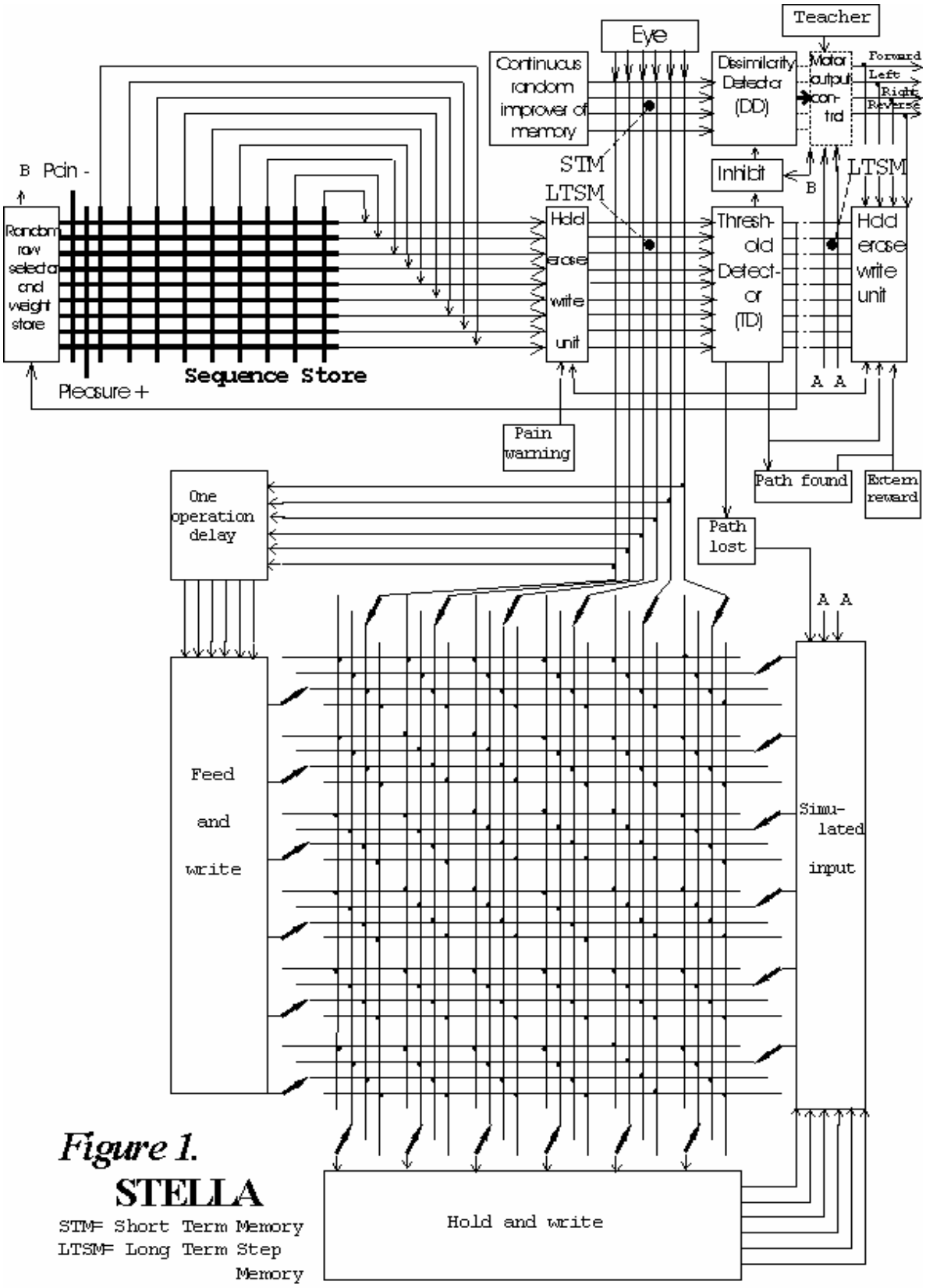

Figure 1.

STELLA

STM= Short Term Memory
LTSM= Long Term Step Memory

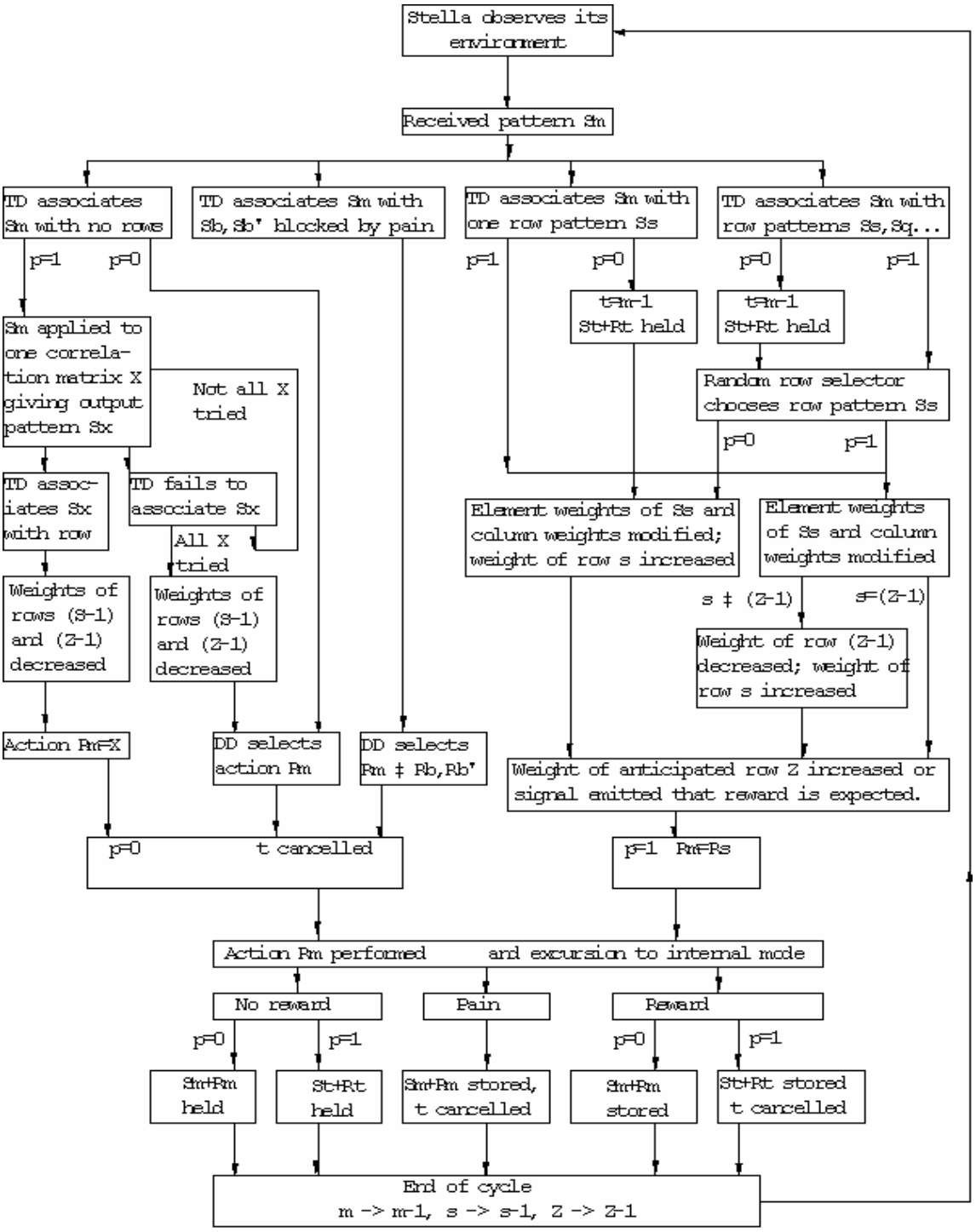
Stella observes its environment
Received pattern Sm
TD associates Sm with no rows
TD associates Sm with Sb, Sb' blocked by pain
TD associates Sm with one row pattern Ss
TD associates Sm with row patterns Ss, Sq...
p=1
p=0
p=1
p=0
p=0
p=1
t=m-1 St+Rt held
t=m-1 St+Rt held
Sm applied to one correlation matrix X giving output pattern Sx
Not all X tried
Random row selector chooses row pattern Ss
p=0
p=1
TD associates Sx with row
TD fails to associate Sx
All X tried
Element weights of Ss and column weights modified; weight of row s increased
Element weights of Ss and column weights modified
s ≠ (Z-1)
s=(Z-1)
Weights of rows (S-1) and (Z-1) decreased
Weights of rows (S-1) and (Z-1) decreased
Weight of row (Z-1) decreased; weight of row s increased
Action Pm=X
DD selects action Pm
DD selects Pm ≠ Pb, Pb'
Weight of anticipated row Z increased or signal emitted that reward is expected.
p=0
t cancelled
p=1 Pm=Ps
Action Pm performed
and excursion to internal mode
No reward
Pain
Reward
p=0
p=1
p=0
p=1
Sm+Pm held
St+Rt held
Sm+Pm stored, t cancelled
Sm+Pm stored
St+Rt stored t cancelled
End of cycle
m -> m-1, s -> s-1, Z -> Z-1

Figure 2